BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Penelitian Terdahulu

Berdasarkan hasil studi literatur, terdapat beberapa penelitian terdahulu yang bisa dijadikan sebagai acuan dalam pengerjaan penelitian ini. Terdapat banyak penelitian yang berhubungan dengan klasifikasi konten *hoax.* Namun meskipun demikian, tentunya terdapat perbedaan pada setiap penelitian. Tabel 1 akan menjelaskan perbedaan penelitian terdahulu yang berkaitan dengan konten *hoax* yang menjadi acuan dalam penelitian ini.

Tabel 1. Penelitian Terdahulu

No	Judul & Tahun	Peneliti	Tujuan	Metode	Hasil
1.	Filtering Hoax menggunakan *Naïve Bayes Classifier* (2018)	P. D Utami, Risna Sari	Mempermudah para pengguna thread forum (female daily) dalam mendeteksi komentar *hoax*	Menggunakan metode *Naive Baayes Classifier*	Ketika data training ada penambahan, yang awalnya sama rata antara *hoax* dan fakta akan menyebabkan nilai probabilitas kelas pada sistem tidak seimbang. Maka setiap kelas data training harus memiliki jumlah seimbang. Dan hasil akurasi yang diperoleh sebesar 88%.
2.	Klasifikasi *Hoax* pada Berita Kesehatan Berbahasa Indonesia	Andre Rino Prasetyo, Indriati, Putra Pandu	Menerapkan metode MKNN untuk mengklasifikasi berita dengan topik kesehatan	Menggunakan metode *Modified K-Nearest Neighbor*	Hasil akurasi yang diperoleh sebesar 75% dengan nilai *k* terbaik 4. Dalam menentukan klasifikasi, konten tiap berita mempengaruhi

	dengan menggunakan *Modified K-Nearest Neighbor* (2018)	Adikara	dalam kategori fakta atau *hoax*		hasil, karena sistem menghitung dari frekuensi kata. Hasil akurasi yang diperoleh tadi tidak terlalu tinggi, karena topik berita kesehatan masih terdapat kata yang kurang baku atau singkatan yang mempersulit dalam mengklasifikasi dengan tepat.
3.	Eksperimen pada Sistem Klasiikasi Beita *Hoax* Berbahasa Indonesia Berbasis Pembelajaran Mesin (2015)	Erissya Rasywir, Ayu Puwaianti	Melakukan percobaan menggunakan beberapa metode untuk mengklasifikasi berita *hoax*	Menggunakan metode *naïve bayes, SVM*, C4.5	Terdapat 2 pengujian, yaitu menggunakan seleksi fitur dan tanpa seleksi fitur. Seleksi fitur yang digunakan yaitu *Information Gain, Mutual Information, Chi-Square, Term Frequency* dan *TDxIDF* yang nantinya akan menggunakan operasi union dan intersection. Dan hasil terbaik dari seleksi fiturnya sendiri adalah fitur unigram, yaitu

					penggunaan operasi union antara *MI* dan *IG*. Menghasilkan akurasi baik menggunakan metode *naïve bayes* sebesar 91,36%
4.	Deteksi Konten *Hoax* Berbahasa Indonesia pada Media Sosial menggunakan Metode *Levenshtein Distance* (2018)	Frista Gifti W	Mendeteksi konten *hoax* dalam sebuah sistem yang menerapkan *levenshtein distance*	*Levenshtein Distance*	Hasil *precision, recall,* dan akurasi menghasilkan nilai yang konsisten. Karena pada scenario pengujian yang ke 2 memiliki data yang lebih banyak dari scenario 1. Jumlah kemunculan kata dan banyaknya dokumen mempengaruhi dalam perhitungan bobot. Penerapan metode *levenshtein distance* sendiri yang dipadukan dengan perhitungan bobot TF IDF mampu mendeteksi berita dengan baik

2.2 Perumusan Hipotesis

Hipotesis merupakan dugaan sementara akan jawaban suatu masalah dalam penelitian dan masih perlu dibuktikan kebenarannya. Penelitian ini dilakukan untuk mengklasifikasi berita *hoax* dengan beragam topik menggunakan metode *Modified K-Nearest Neighbor*. Berikut merupakan hipotesis dari penelitian ini:

H_0 : Metode *Modified K-Nearest Neighbor*, dalam penerapannya mampu melakukan klasifikasi berita *hoax* dengan beragam topik.

2.3 Landasan Teori

2.3.1 Hoax

Menurut Kamus Besar Bahasa Indonesia(KBBI) kata *hoax* bermakna berita bohong. Sehingga *hoax* merupakan berita yang berisikan informasi bohong. Bohong dalam artian bisa berisikan informasi yang palsu maupun informasi belum pasti bahkan bukan sebuah fakta mengenai kebenarannya. Kebanyakan berita bohong membuat para pembacanya mudah mempercayai dan dengan mudahnya menyebarluaskan berita bohong tersebut.

Hoax memang sudah menyebar dikalangan masyarakat. Berawal dari para pembuat berita yang berisikan opini, data foto maupun gambar yang bersifat *hoax* dalam membagikan menggunakan media sosial yang ada[6]. Di Indonesia sendiri memiliki undang-undang yang mengatur mengenai penyebaran berita bohong, yaitu Undang-Undang No 11 tahun 2008 tentang Informasi dan Transaksi Elektronik Pasal 28 Ayat 1.

2.3.2 Text Mining

Text mining merupakan solusi yang dapat digunakan dalam segala hal masalah yang berhubungan dengan text yang berjumlah besar. *Text* mining sendiri bisa diartikan suatu proses penemuan informasi teranyar yang tiada terkuak sebelumnya dengan melakukan proses dan analisa data dengan jumlah besar. Dalam melakukan analisa suatu *text* tidak terstruktur, *text mining* mencoba untuk mengolah informasi dari 1 *text* ke *text* lainnya berlandaskan aturan khusus. Dengan harapan hasil yang

diperolah adalah sebuah informasi baru yang sebelumnya tidak terkuak[7].

Sama halnya dengan data mining, text mining juga mempunyai beberapa masalah seperti jumlah data besar, dimensi yang tinggi, data terus berubah dan data noise. Dalam text mining juga memiliki tujuan dan menggunakan proses yang sama dengan data mining. Namun dalam text mining data dalam bentuk data tidak terstruktur atau paling tidak semi struktur dan text[7].

2.3.3 Text Preprocessing

Dalam proses klasifikasi text, perlu dilakukan tahap pertama yaitu pengolahan data atau disebut dengan *Text Preprocessing.* Dimana melakukan seleksi data yang diproses pada setiap dokumen yang akan melewati beberapa tahapan yang disesuaikan dengan kebutuhan penelitian ini, yaitu sebagai berikut:

a. Tokenizing

Tahapan dimana merubah atau penguraian dari kalimat yang berupa paragraf menjadi tiap kata. Dengan tujuan agar dapat dilakukan perhitungan bobot dari setiap kata yang muncul.

b. Stopword Removal

Stopword Removal merupakan tahapan penghapusan kata tidak relevan atau tidak penting berdasarkan list data *stopword* yang telah disiapkan. Sehingga, ketika terdapat kata yang tertera dalam *list stopword.* Maka akan dihapus kata tesebut.

c. Stemming

Tahapan dimana mengubah kata yang ada di dalam tiap paragraph menjadi kata dasarnya. Guna menghilangkan kata awalan dan imbuhan pada setiap katanya.

2.3.4 Pembobotan Kata

TF-IDF merupakan salah satu algoritma pembobotan kata yang sering digunakan. Perhitungan pembobotan kata ini diperlukan untuk mengetahui seberapa penting kata tersebut. Sehingga kata yang memiliki bobot tinggi semakin penting pula kata tersebut. Untuk

pengertian *Term Frequency* ini sendiri merupakan jumlah berapa kali kata tersebut muncul dalam dokumen. *Insverse Document Frequency* untuk menghitung probabilitas penemuan kata dalam teks[8].

$$W_{ij} = tf_{ij} * log \frac{N}{df_i} \quad (1)$$

Keterangan:

W_{ij} = bobot kata *i* dalam dokumen *j*

tf_{ij} = *frequency* dari kata *i* dalam dokumen *j*

N = jumlah dokumen

df = jumlah dokumen yang mengandung kata

2.3.5 Algoritma KNN

Metode *K-NearestNeighbor* merupakan salah satu metode yang sangat mudah dan sering digunakan dalam penerapan klasifikasi. Selain itu, metode *K-Nearest Neighbor* ini mampu dalam menangani noise data, tidak rumit dan bisa digunakan dalam komputasi yang sangat besar[9]. *K-Nearest Neighbor* bertujuan mengklasifikasi objek baru berdasarkan jarak yang paling terdekat[10]. Berikut merupakan langkah metode *K-Nearest Neighbor*:

1. Menghitung jarak *Eucledian*.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(ar(xi) - ar(xj))^2} \quad (2)$$

Keterangan:

$d(xi,xj)$ = jarak *Eucledian*

x_i = *record* ke-*i*

x_j = *record* ke-*j*

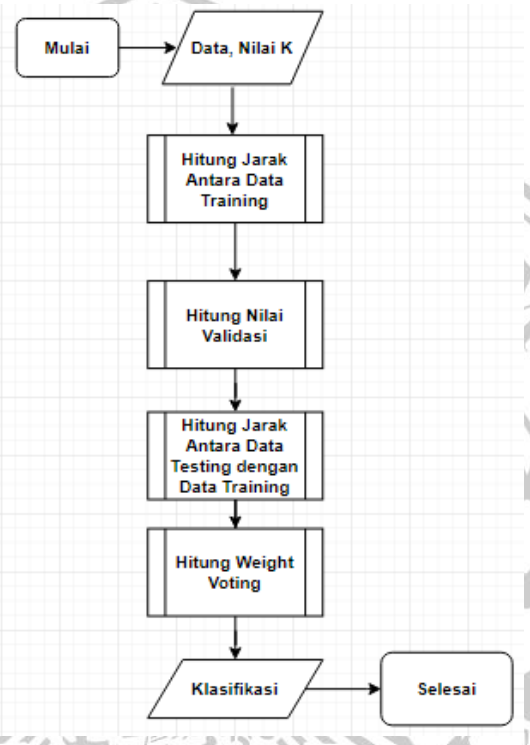
a_r = data ke-*r*

2. Hasil dari *Eucledian Distance* diurutkan.
3. Mengambil hasil klasifikasi terdekat berdasarkan nilai K.
4. Hasil target berdasarkan kelas terbanyak.

2.3.6 Algoritma MKNN

Modified K-Nearest Neighbor Neighbor merupakan pengembangan dari algoritma *K-Nearest Neighbor*. Di dalam metode

modified k-nearest neighbor ini ada proses penambahan, yaitu proses perhitungan nilai validitas dan perhitungan *weight voting*[11]. *Modified K-Nearest Neighbor* ini diciptakan untuk meningkatkan nilai akurasi *K-Nearest Neighbor* berdasarkan penambahan perhitungan tersebut[9]. Berikut merupakan diagram alir metode *Modified K-Nearest Neighbor*.



Gambar 1. Diagram Alir Metode MK-NN

1. Jarak Eucledian

Tahap perhitungan *eucledian distance* ini terdapat 2 tahapan. Tahap pertama merupakan perhitungan nilai jarak *eucledian* antar data *training*. Selanjutnya adalah perhitungan antar data *training* dengan *data testing*[11].

$$d\,(xi, yi) = \sqrt{\sum_{i=0}^{n}(xi - yi)^2} \quad (3)$$

Keterangan:

d = jarak
x = data *training*
y = data *testing* yang akan diklasifikasikan

2. Nilai validitas

Hal pertama dalam perhitungan, data *training* harus tervalidasi dan validasi tergantung pada tetangganya. Semua data *training* harus melalui proses validasi. Hasil dari nilai validitas tersebut akan digunakan sebagai informasi data[11].

$$Validity(x) = \frac{1}{H}\sum_{i=0}^{H} S(lbl(x), lbl(Ni(x))) \quad (4)$$

Keterangan:

H = jumlah titik terdekat
lbl(x) = kelas x
lbl(Ni(x)) = label kelas terdekat x

Perhitungan antara titik x dan data ke-i dari tetangga terdekat merupakan kegunaan dari fungsi S. berikut merupakan persamaan dari fungsi S[11].

$$S(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \quad (5)$$

Keterangan:

a = kelas a pada data *training*
b = kelas lain selain a pada data *training*

2. Weight Voting

Perhitungan dalam *weight* tiap tetangga dilakukan dengan perhitungan 1/(de + 0,5). Dilanjutkan dengan nilai validitas tiap data *training* akan dilakukan perkalian dengan *weight* yang berdasarkan jarak *eucledian*[11].

$$W(i) = Validity\ (i)\ x\ \frac{1}{de + 0.5} \quad (6)$$

Keterangan:

W(i) = perhitungan *weight voting*
Validity(i) = nilai validitas
de = jarak *eucledian*

2.3.7 Pengujian Klasifikasi

Proses pengujian dalam sistem sangatlah penting. Karena dalam perhitungan pengujian akan menunjukkan seberapa berhasilnya sistem ini dibuat yang dapat dilihat dari hasil nilai keakurasiannya. *Confusion Matrix* merupakan metode pengujian untuk melihat tingkat nilai keakurasian suatu teknik klasifikasi. Menggunakan temu kembali tak berperingkat[4].

Tabel 2. Tabel Confusion Matrix

Hasil Prediksi	Hasil Aktual	
	Fakta	*Hoax*
Fakta	TP	FP
Hoax	FN	TN

Keterangan:

- (TP) *True Positive* adalah jumlah data positif yang terprediksi benar positif
- (TN) *True Negative* adalah jumlah data negatif yang terprediksi benar negatif
- (FP) *False Positive* adalah jumlah data negatif namun terprediksi salah positif
- (FN) *False Negative* adalah jumlah data positif namun terprediksi salah negatif

Berikut merupakan rumus *precision*, *recall*, *f-measure* dan akurasi:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F\text{-}Measure = \frac{2 \text{ x } P \text{ x } R}{P+R} \quad (9)$$

$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+TN} \quad (10)$$